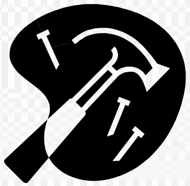
The Comparison of Two Measurement Devices

2006 Joint Research Conference
June 7–9, 2006

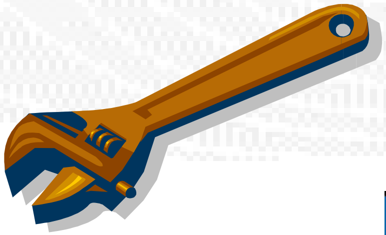
Joseph G. Voelkel (Bruce Siskowski
CQAS, COE, RIT Reichert, Inc.)

Topics



- Problem, Example, Mathematical Model
- Comparison: Regression? Bland-Altman?
- More Models, Identifiability Problem, Bland-Altman
- A Richer Data Set and a Larger Model

- Comparison to Gage R&R
- Mandel's Estimates
- Data Analysis
  - Informal—Graphs, Background Assumptions
  - Formal—Likelihood Methods



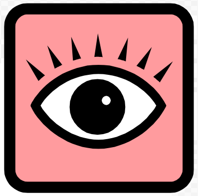
The Problem

- Two measuring devices need to be compared
  - Say, new vs old
  - (Can extend to more than two…)
- No Standard
  - No standard exists for what is the right answer
  - A standard exists but is hard to come by – $$
  - A standard exists but is not realistic

The Problem

- Examples
  - Blood pressure
  - Cardiac Output
    - Fick method
    - Dyel dilution
    - Thermal dilution
- “Correct” answer hard to come by
- “Gold Standard” also has measurement error

Tonometer Example



- Medical screening device that measures intra-ocular pressure of the human eye.
- Pressure acts on retina and optic nerve.
- Increased sustained pressures above 23mm Hg can lead to vision loss condition—glaucoma.
- If tonometer indicates possible risk, an M.D. of ophthalmology runs other detailed tests for a more accurate diagnosis.

Tonometer Example

- Problem with tonometer calibration
  - Difficult to put pressure sensors inside the human eye (!) to measure “exact” values
  - Sensor insertion surgery exists but would change the eye anyway…
- Original gold standard is Goldman Applanation Tonometer (GAT) that touches the eye
- Example of a contact tonometer

Reichert

Tonometer Example

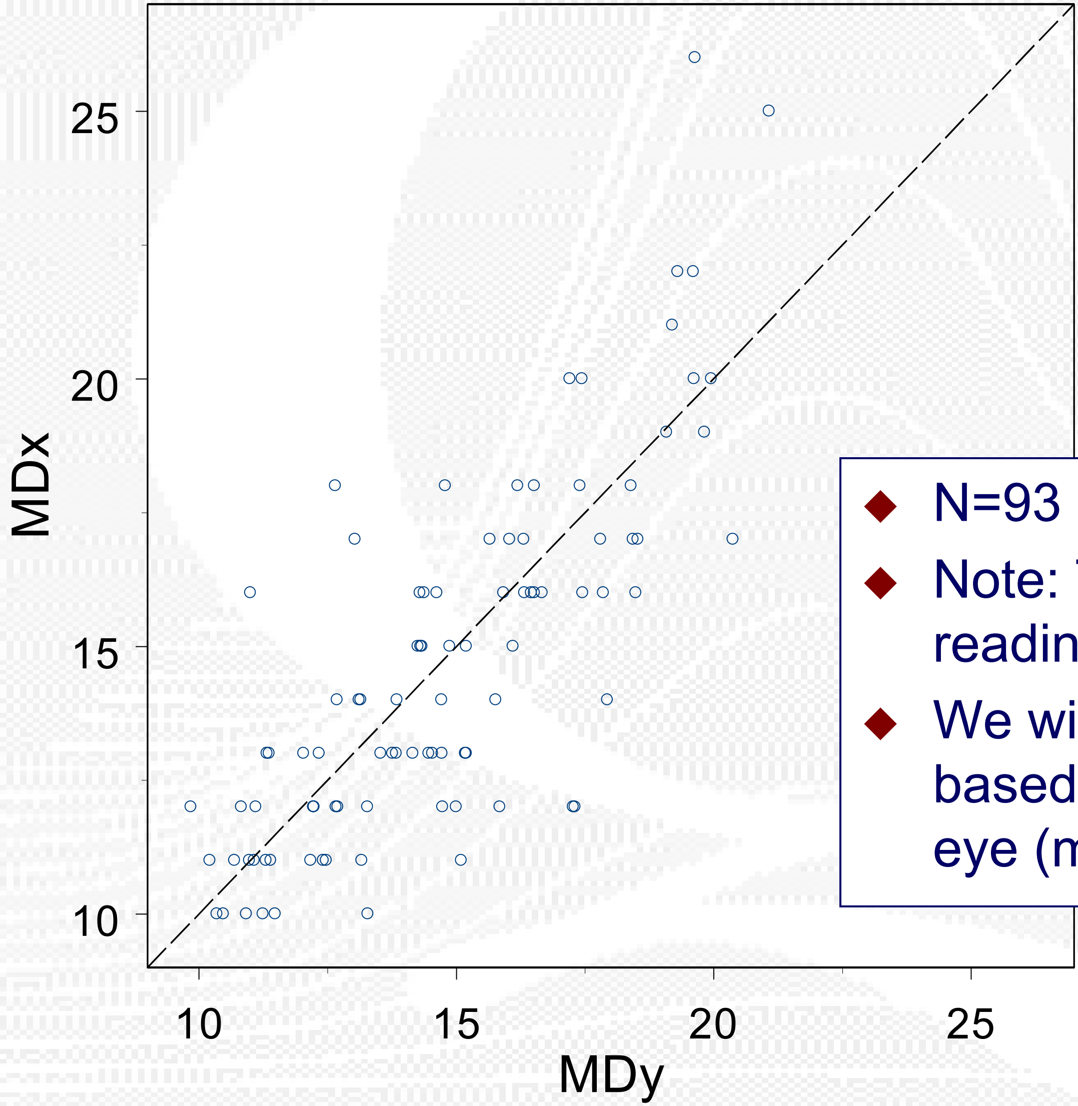
- Reichert invented several non-contact air-puff versions since 1972 that
  - Do not require eye anesthetic drops
  - Do reduce operator variation via computerized automation.
- Reichert's goal is to employ better statistical tests to see if Reichert tonometers have *less measurement repeatability variation* than the GAT
- Most popular technique (Bland-Altman) only checks "*agreement*" and *bias* (more to follow)

Reichert

Tonometer Example

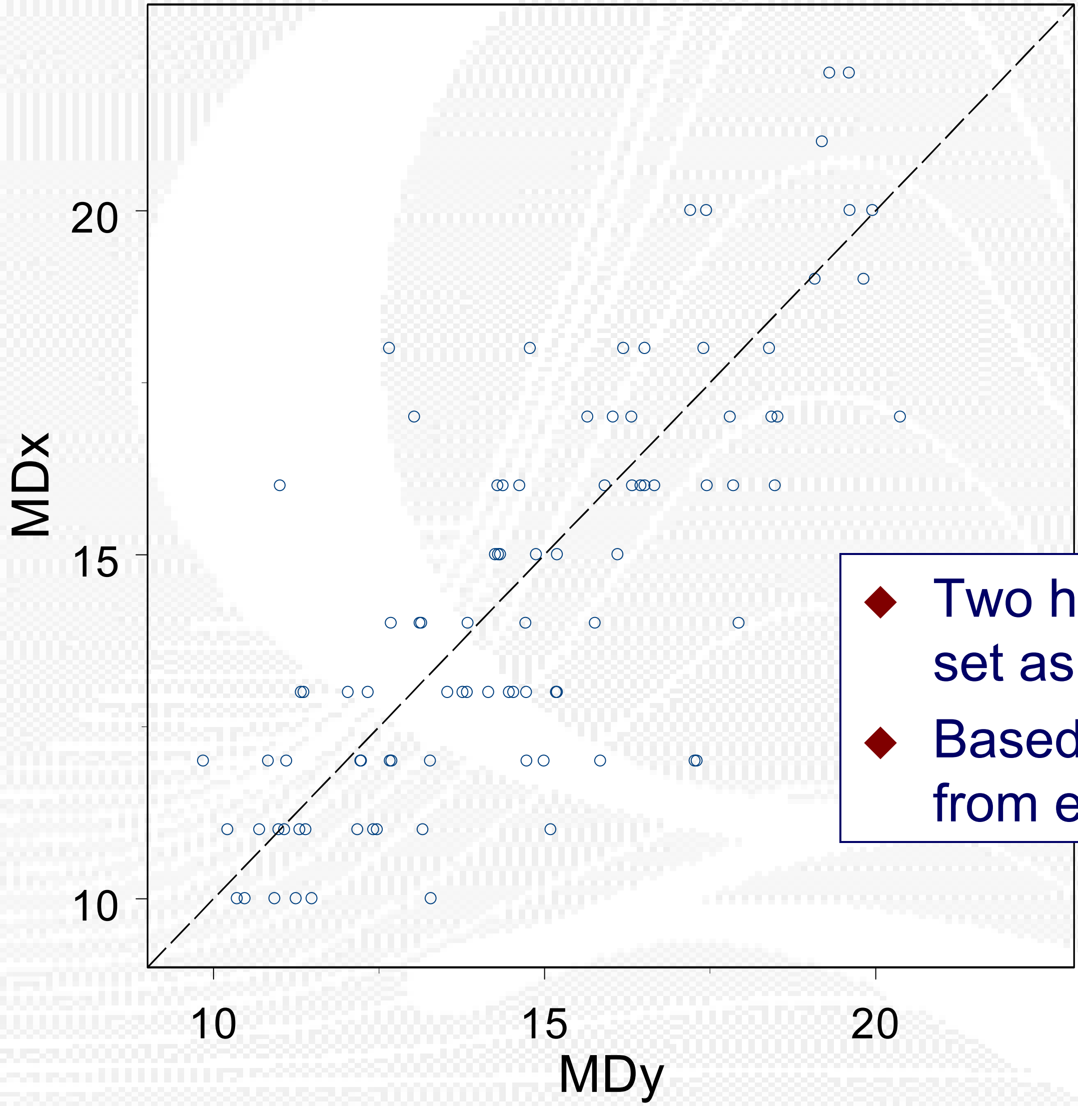
- Two tonometers (different models). The reference device is called MD_x and the device under test is MD_y
- Example slightly simplified from original study. Only measurements of the left eye, in mm Hg. (Coded.)
- Study performed by selecting a sample of subjects. Each subject measured with MD_x and then with MD_y

Data



- N=93
- Note: This is only based on one reading from each eye
- We will later consider averages based on multiple readings per eye (more common)

Data



- Two highest MD_x values set aside → N=91
- Based on only one reading from each eye

Are the Two Devices Equivalent? And Other Questions…

- What does it mean to say “equivalent”?
- And if they are not equivalent, in what way are they not equivalent?

A (Tentative) Mathematical Model

MD_x $\quad x_1 \quad x_2 \quad x_3 \quad x_4 \; \ldots \; x_N$ Long-term average *right now* ("true?")

$X_1 \quad X_2 \quad X_3 \quad X_4 \; \ldots \; X_N$ Observed

- What does it mean to say “equivalent”?

MD_y $\quad y_1 \quad y_2 \quad y_3 \quad y_4 \; \cdots \; y_N$

$Y_1 \quad Y_2 \quad Y_3 \quad Y_4 \; \ldots \; Y_N$

A Mathematical Model

1. Where did these subjects come from??

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad \ldots \quad x_N$

r.s. size N from a pop'n

2. What do the x_i's look like in the population? $x_i \sim \text{ind } N\left(\mu_x, \sigma_x^2\right)$

◆ Our x_i's…

A Mathematical Model

$$x_i \sim \text{ind } N\left(\mu_x, \sigma_x^2\right)$$

3. What do we observe?

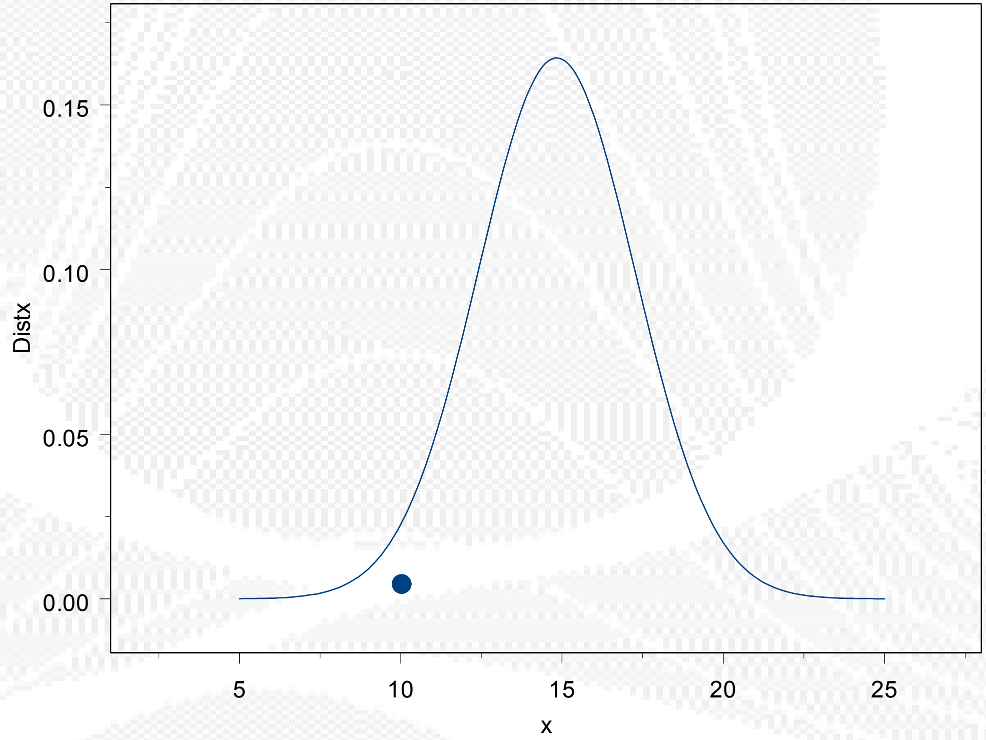
$$X_i = x_i + e_i, \quad e_i \sim \text{ind } N\left(0, \sigma_e^2\right)$$

$\boxed{e_i \text{ is the } x_i \textit{ measurement error}}$

A Mathematical Model

- The x distribution and, say, x_1

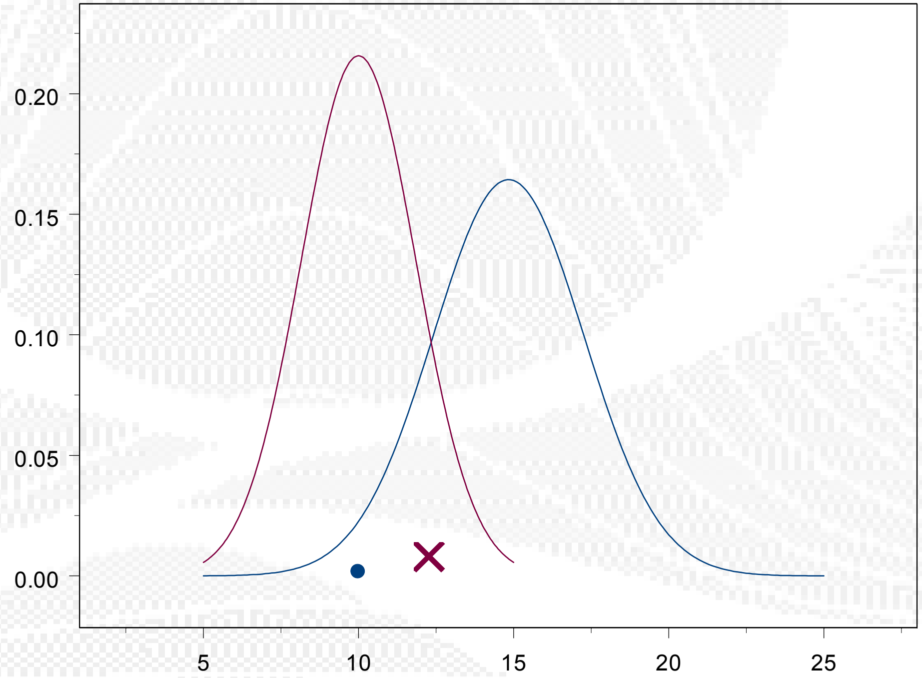
$$x_i \sim \text{ind } N\left(\mu_x, \sigma_x^2\right)$$



A Mathematical Model

- The x distribution and, say, x_1
- The X distribution at x_1. Also, X_1

$$X_i = x_i + e_i, \quad e_i \sim \text{ind } N\left(0, \sigma_e^2\right)$$



A Mathematical Model, under Equivalency

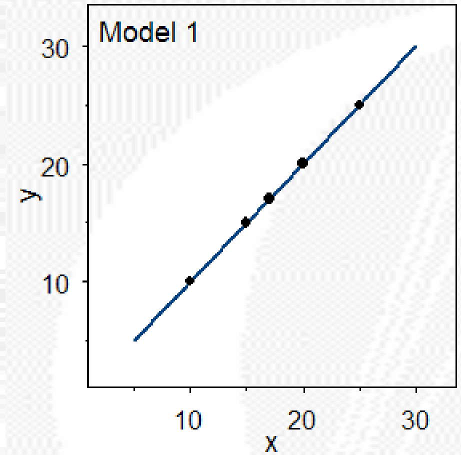
4. What about the y_i's?

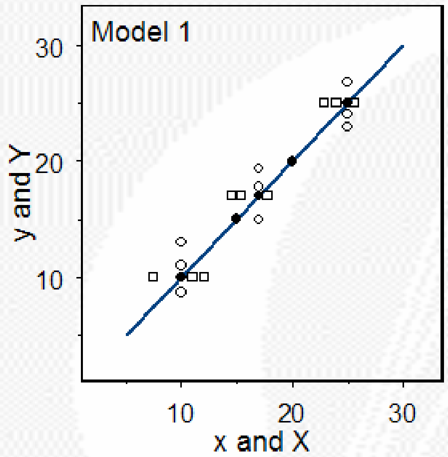
- Should have some connection to the x_i's…
- Equivalency $\equiv$ **Model 1**

$$y_i = x_i$$

$$Y_i = y_i + u_i, \quad u_i \sim \text{ind}\, N\left(0, \sigma_u^2\right)$$

$$\sigma_u^2 = \sigma_e^2$$





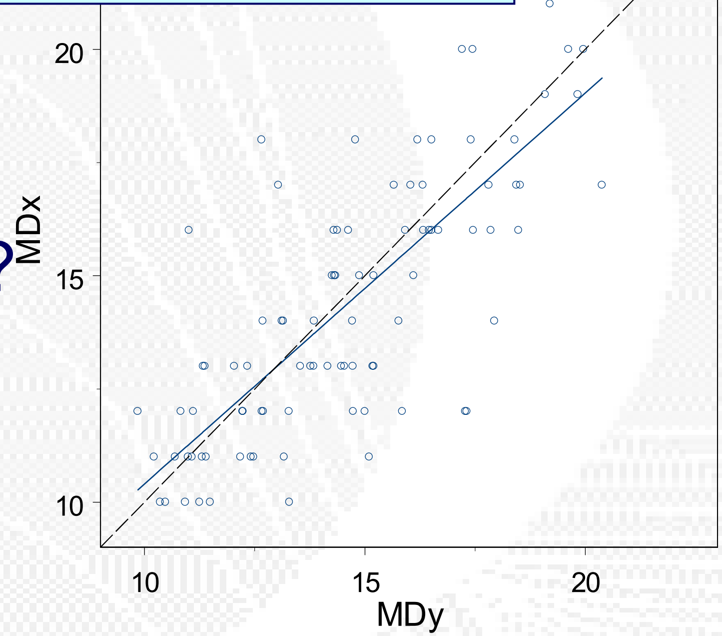
Topics

- Problem, Example, Mathematical Model
- Comparison: Regression? Bland-Altman?
- More Models. Identifiability Problem, Bland-Altman
- A Richer Data Set and a Larger Model
- Comparison to Gage R&R
- Mandel’s Estimates
- Data Analysis
  - Informal—Graphs, Background Assumptions
  - Formal—Likelihood Methods

Regression? Gap in Theory vs Practice

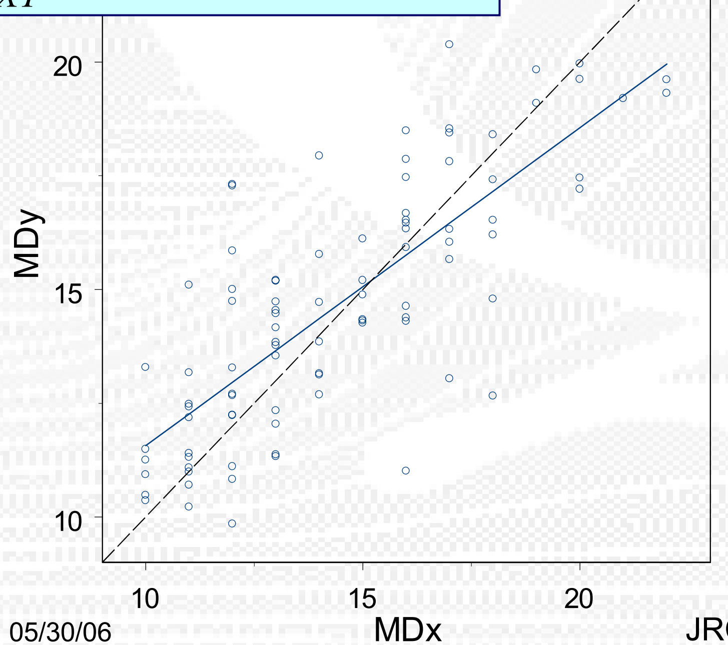
- Medical researchers
  - Regression of X on Y?
  - Regression of Y on X?
  - Correlation of Y and X?

$$\hat{\beta}_{XY} = 0.86 \quad \text{s.e.}=0.07$$



$$\hat{\rho}_{XY} = 0.78$$

$$\hat{\beta}_{XY} = 0.70 \quad \text{s.e.}=0.06$$



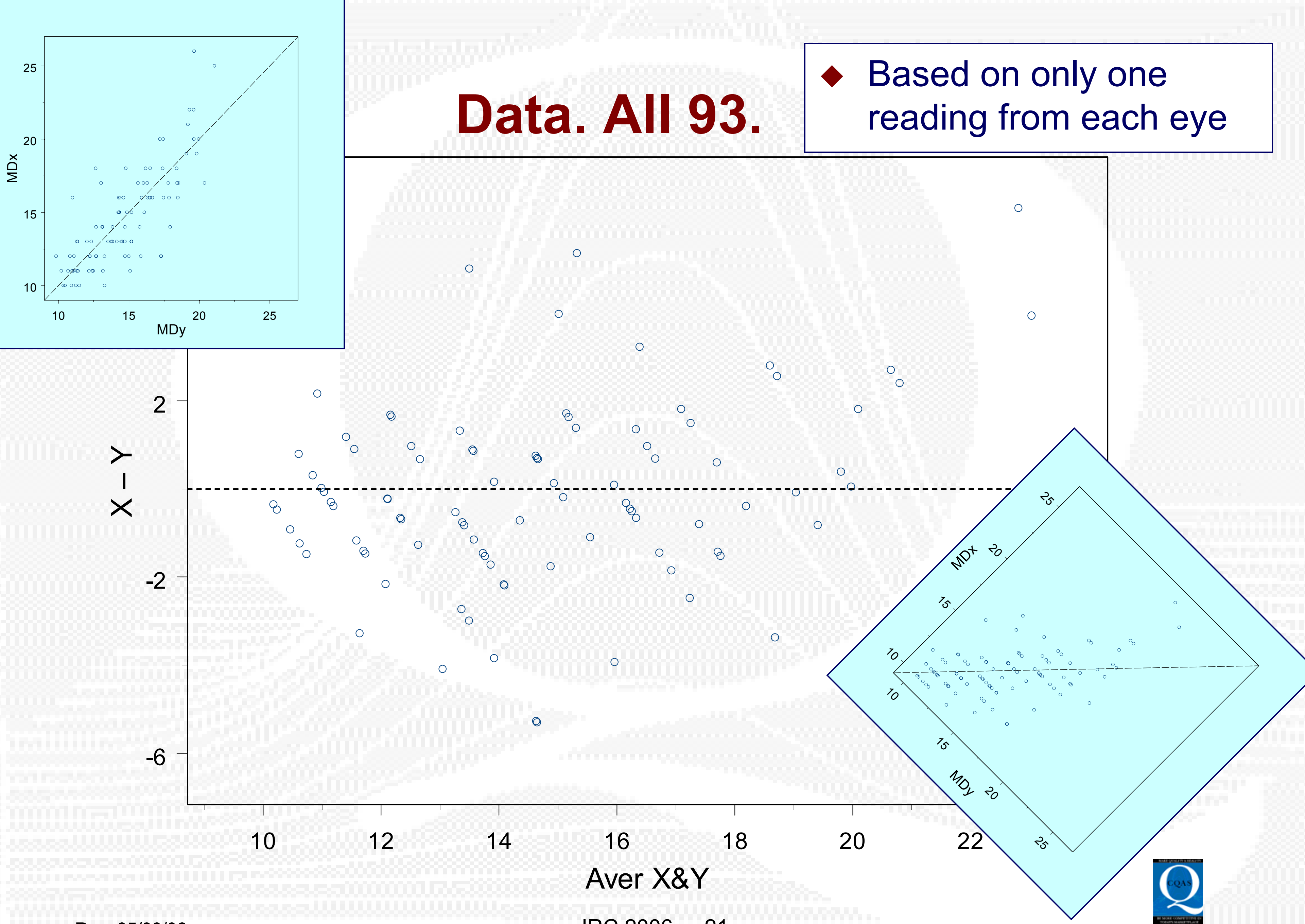
- Based on only one reading from each eye

Bland-Altman

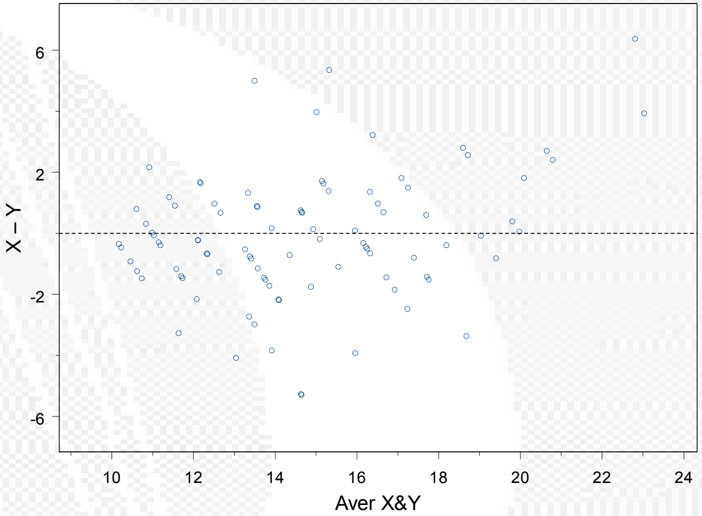
- Instead of Y vs X…
- Plot Y-X vs average(Y & X)
  - An example of a difference-mean plot
- Then look for *agreement*
- Very popular. One of the 10 most highly cited papers in statistics.

Data. All 93.

- Based on only one reading from each eye



Bland-Altman

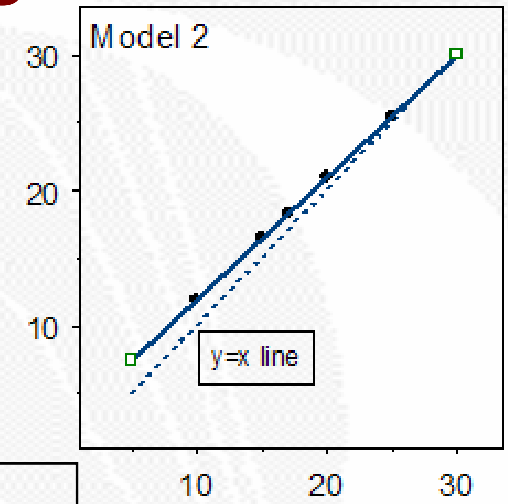


- Use graph to check for
  - Outliers
  - Linear trends, bias
  - More Spread at higher Aver(X&Y) values
    - If so, try log transformation
- If all OK, summarize agreement by s.e.(X–Y)
- Here, if only use N=91, get s.e.=2.0
- Bland-Altman has become a standard method, accepted way to make comparisons

Topics

- Problem, Example, Mathematical Model
- Comparisons: Regression? Bland-Altman?
- More Models. Identifiability Problem, Bland-Altman
- A Richer Data Set and a Larger Model
- Comparison to Gage R&R
- Mandel's Estimates
- Data Analysis
  - Informal—Graphs, Background Assumptions
  - Formal—Likelihood Methods

Back to Model Thinking…



- So far, have just defined "equivalent" devices.
- More generally, consider model with possible *linear bias*

Model 2

$$x_i \sim \text{ind } N\left(\mu_x, \sigma_x^2\right)$$

$$y_i = \beta_0 + \beta_1 x_i$$

$$X_i = x_i + e_i,\ e_i \sim \text{ind } N\left(0, \sigma_e^2\right)$$

$$Y_i = y_i + u_i,\ u_i \sim \text{ind } N\left(0, \sigma_u^2\right)$$

$$\sigma_u^2 = \sigma_e^2$$

Another Model…

- Last model—*possible linear bias* but *same measurement s.d.'s*
- This model—*no linear bias* but *possible different measurement s.d.'s*

Model 2'

$$x_i \sim \text{ind } N\left(\mu_x, \sigma_x^2\right)$$
$$y_i = x_i$$
$$X_i = x_i + e_i,\ e_i \sim \text{ind } N\left(0, \sigma_e^2\right)$$
$$Y_i = y_i + u_i,\ u_i \sim \text{ind } N\left(0, \sigma_u^2\right)$$

And Another Model…

- A model with possible *linear bias* and *different measurement s.d.'s*

Model 3

$$x_i \sim \text{ind } N\left(\mu_x, \sigma_x^2\right), \quad y_i = \beta_0 + \beta_1 x_i$$
$$X_i = x_i + e_i,\ e_i \sim \text{ind } N\left(0, \sigma_e^2\right)$$
$$Y_i = y_i + u_i,\ u_i \sim \text{ind } N\left(0, \sigma_u^2\right)$$

- Very reasonable. MD_x and MD_y measuring the same feature, but possibly un-calibrated and possibly with different precision.
- Models 1-3: "structural," "measurement-error," models (Fuller (1987))

Literature …

- Vast literature on this and related problems
  - Lord (1960), Grubbs (1948), Pearson (1902); Thompson (1963), Jaech (…)
    - Estimating var's in instruments w/o repeats
  - Wald (1940), Geary (1949), Tukey (1951)
    - Use of add'l info: Instrumental variables
  - Mandel (1959), Cochran (1968)
    - Interlab comparison; survey examples.
  - Lindley (1947), Neyman (1951), Kendall (1951), Wolfowitz (1952), Madansky (1959), Berkson (1950), Box (1961)

Information in the Data for Model 3?

- Under Model 3 assumptions, it is well known that the minimal sufficient statistic is 5 –dimensional:

$$\boxed{\bar{X}, \bar{Y}, s_X^2, s_Y^2, r_{X,Y} \text{ (or } \widehat{Cov}(X,Y))}$$

$$\boxed{\mu_x, \sigma_x^2, \beta_0, \beta_1, \sigma_e^2, \sigma_u^2}$$

$$x_i \sim \text{ind } N\left(\mu_x, \sigma_x^2\right)$$
$$y_i = \beta_0 + \beta_1 x_i$$
$$X_i = x_i + e_i,\ e_i \sim \text{ind } N\left(0, \sigma_e^2\right)$$
$$Y_i = y_i + u_i,\ u_i \sim \text{ind } N\left(0, \sigma_u^2\right)$$

- However, there are 6 parameters that must be estimated in the Model
- *Unidentifiable* with the data available

Model 3 Problem

- Model 3: *unidentifiable* with the data available
- Bland and Altman still advocate their method…
- Problems with Bland-Altman:
  - *Does not* allow bias to be estimated cleanly
  - *Does not* give a pure estimated measure of agreement, but *does* give a upper bound of it.

$$\boxed{E\left[s_{X-Y}^2\right]=\sigma_x^2\left(\beta_1-1\right)^2+\sigma_e^2+\sigma_u^2}$$

    So, the s.e.=2.0 is a upper bound estimate of the s.d. of the differences

  - *Does not* provide *any* information on relative precision.

Model 3 Problem: Normality?

- Reiersøl (1950)

If e_i and $u_i \sim$ i.i.d Normal, then (β_0, β_1) non-identifiable iff (X_i, Y_i) are constants or bivariate Normal

- Model 3

$$x_i \sim \text{ind } N\left(\mu_x, \sigma_x^2\right)$$
$$y_i = \beta_0 + \beta_1 x_i$$
$$X_i = x_i + e_i,\ e_i \sim \text{ind } N\left(0, \sigma_e^2\right)$$
$$Y_i = y_i + u_i,\ u_i \sim \text{ind } N\left(0, \sigma_u^2\right)$$

- Mostly of theoretical interest

Bland and Altman: A Question

- Is *agreement* really want we want to examine?
- If there is lack of agreement, do we know
  - why?
  - which device, if either, is better?
- No. For example:
  - If the "gold standard" does not agree with the new device, it may be that the new device is very precise and the gold standard is highly variable.

Topics

- Problem, Example, Mathematical Model
- Comparisons: Regression? Bland-Altman?
- More Models. Identifiability Problem, Bland-Altman
- A Richer Data Set and a Larger Model
- Comparison to Gage R&R
- Mandel's Estimates
- Data Analysis
  - Informal—Graphs, Background Assumptions
  - Formal—Likelihood Methods

A Richer Data Set

- If possible, collect more than one observation for each subject.
- Note
  - Bland and Altman advocate this on paper, but most of their examples use one-observation-per-subject for each device (even if more than one observation was available)
  - In any event, they still continue to use the notion of *agreement*

A Richer Data Set

$$
\begin{array}{llllll}
MD_x & x_1 & x_2 & x_3 & x_4 \ \dots & x_N \\
 & X_{11} & X_{12} & X_{13} & X_{14} \dots & X_{1N} \\
 & X_{21} & X_{22} & X_{23} & X_{24} \dots & X_{2N} \\
 & X_{31} & X_{32} & X_{33} & X_{34} \dots & X_{3N}
\end{array}
$$

(Total MD_x data)

Long-term average *right now*

Observed

$$x_i, X_{ji}, i = 1, \dots, N, j = 1, \dots, J$$

A Richer Data Set

- The additional information

$$\begin{array}{ccccc} X_{11} & X_{12} & X_{13} & X_{14} \ldots & X_{1N} \\ X_{21} & X_{22} & X_{23} & X_{24} \ldots & X_{2N} \\ X_{31} & X_{32} & X_{33} & X_{34} \ldots & X_{3N} \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ s_{e1}^2 & s_{e2}^2 & s_{e3}^2 & s_{e4}^2 & s_{eN}^2 \end{array} \Rightarrow s_e^2$$

and s_u^2 for MD_y

- Now: 7 summaries to estimate 6 parameters.

CQAS

A Larger Model

- With 7 summaries to estimate 6 parameters, consider a larger, possibly more realistic, model

Model 4

- What if the two measuring devices are not quite measuring the same feature?

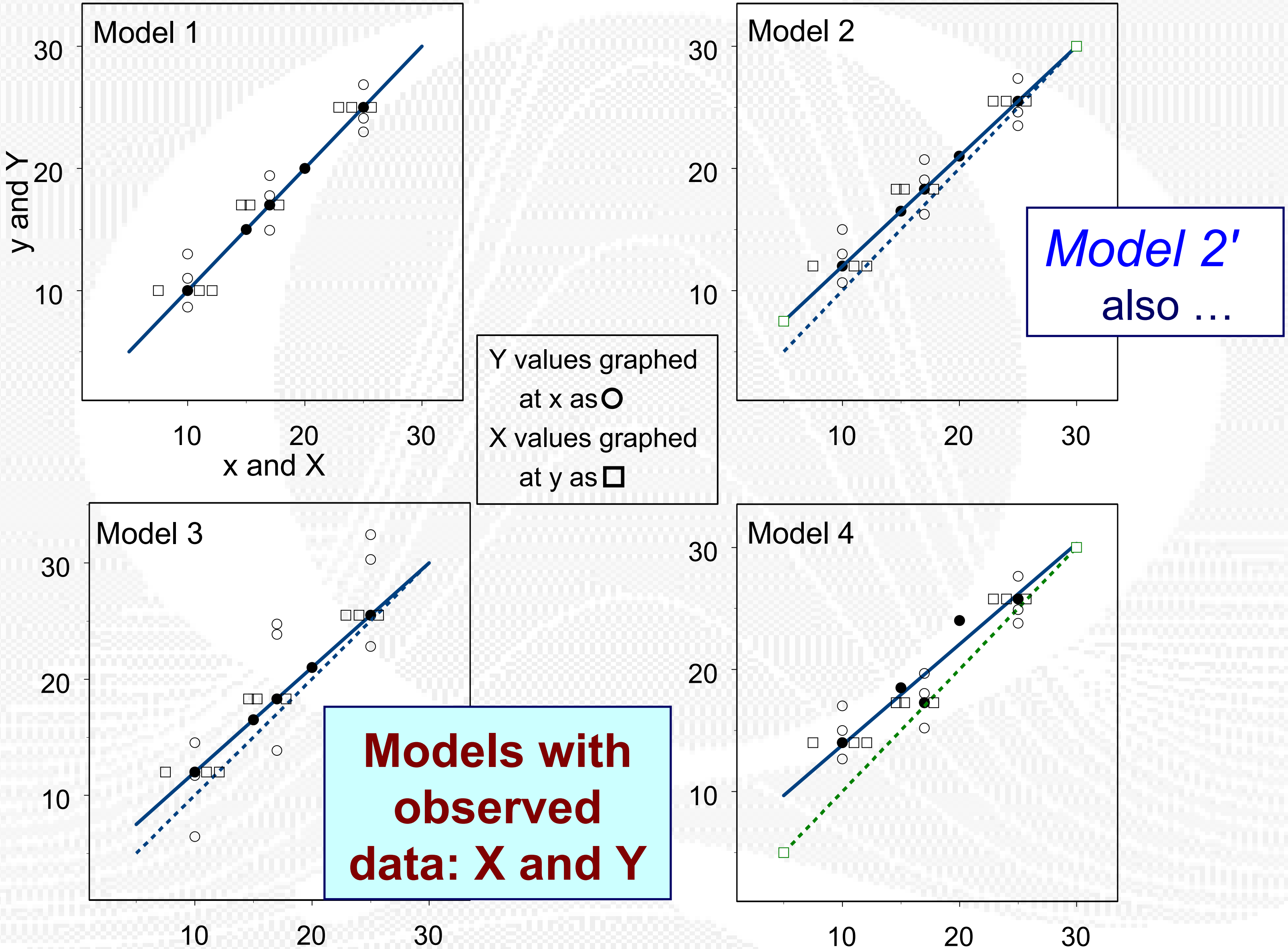
$$x_i \sim \text{ind } N\left(\mu_x, \sigma_x^2\right)$$

$$y_i = \beta_0 + \beta_1 x_i + \delta_i,\ \delta_i \sim \text{ind } N\left(0, \sigma_\delta^2\right)$$

$$X_{ji} = x_i + e_{ji},\ e_{ji} \sim \text{ind } N\left(0, \sigma_e^2\right)$$

$$Y_{ji} = y_i + u_{ji},\ u_{ji} \sim \text{ind } N\left(0, \sigma_u^2\right)$$

- Model 4: structural, but not measurement-error, model.
- Still symmetric in (x,y), but "a problem model"

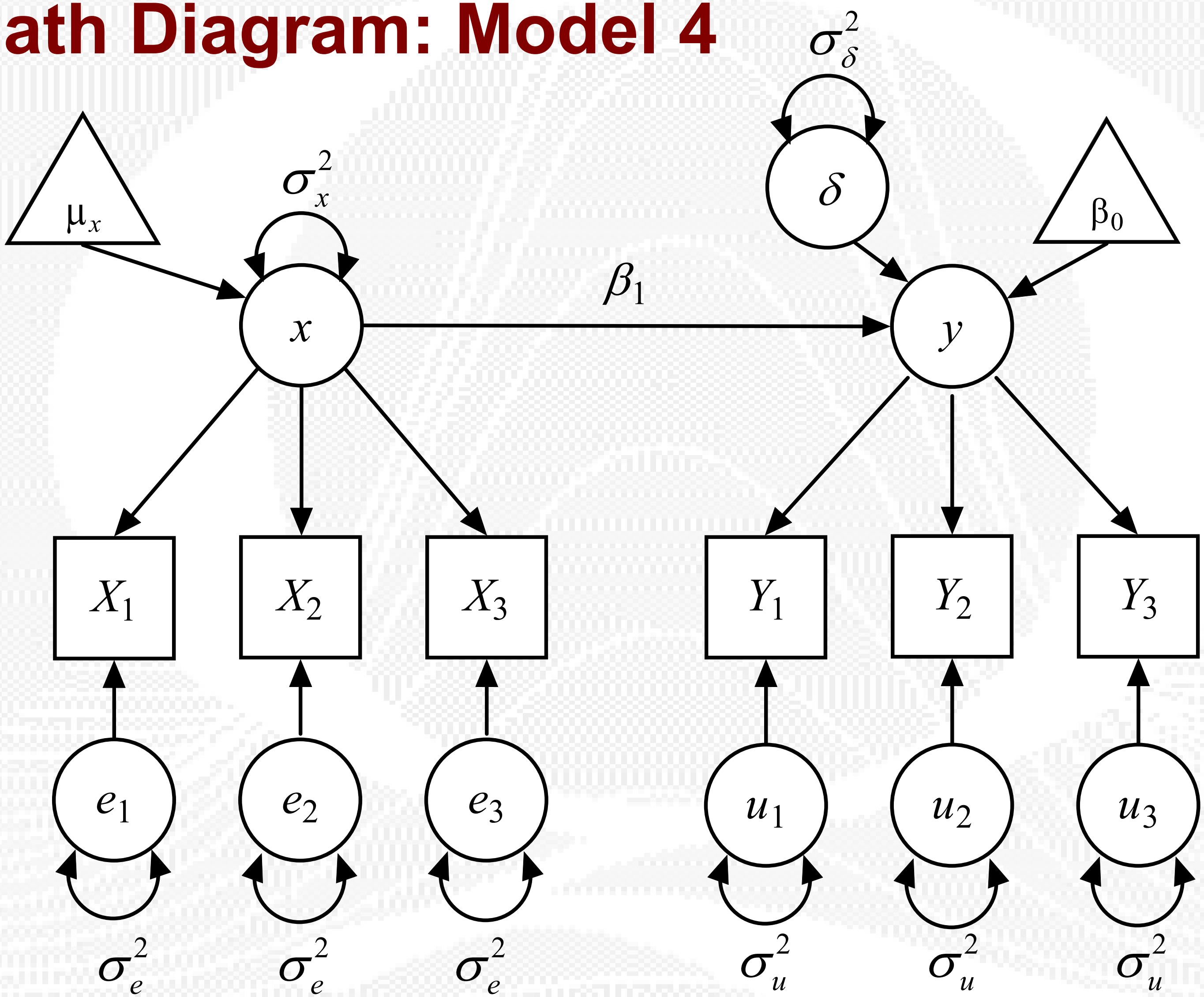


Models with observed data: X and Y

Aside: Path Diagrams

- Common in the sociological literature, e.g. Bollen (1989)
- Unobserved variables (x, y): *latent variables*
  - Intelligence, socio-economic status
- Observed variables (X, Y): *manifest variables*.
  - Scores on IQ test, annual income

Path Diagram: Model 4



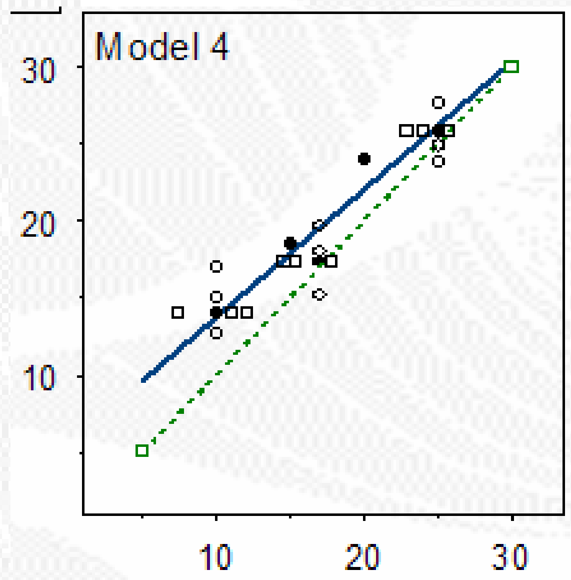
CQAS

Topics

- Problem, Example, Mathematical Model
- Comparisons: Regression? Bland-Altman?
- More Models. Identifiability Problem, Bland-Altman
- A Richer Data Set and a Larger Model
- Comparison to Gage R&R
- Mandel’s Estimates
- Data Analysis
  - Informal—Graphs, Background Assumptions
  - Formal—Likelihood Methods

Comparison to Gage R&R

- One device, several (say two) operators → Two devices
- So, operators as devices…
- General operator differences (vs. specific—linear trend—differences & deviations from it)



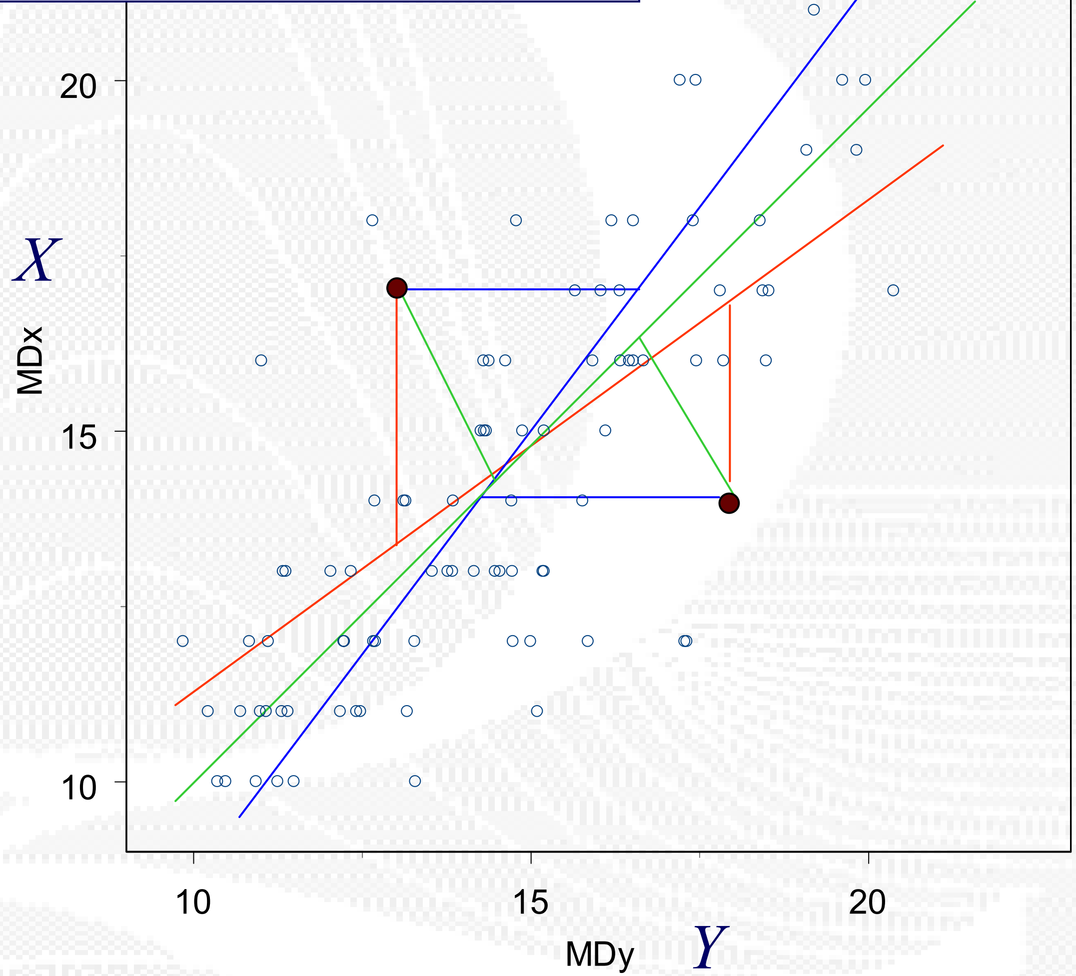
- In usual case, assumes each operator's measurement error equal (vs. looking for different device precision)
- Often, small study (10 parts…), with poor estimates (vs. more data & better estimates)

Mandel's Estimate and The Regression Problem

- Mandel (1984) considered Model 3 (possibly un-calibrated and different precision, but measuring same feature)
- He noted a rule for finding the best fitting line (estimating the relation between x and y, not X and Y)
- (A rediscovery? Lindley (1947))

- Based on only one reading from each eye

- All meas't error in X:
  →Least Squares based on Regression of X on Y
- All meas't error in Y:
  →Least Squares based on Regression of Y on X
- Equal meas't error in X & Y: → Least Squares based on 45° line
- General Case: Least Squares based on *k*° line



Topics

- Problem, Example, Mathematical Model
- Comparisons: Regression? Bland-Altman?
- More Models. Identifiability Problem, Bland-Altman
- A Richer Data Set and a Larger Model
- Comparison to Gage R&R
- Mandel’s Estimates
- Data Analysis
  - Informal—Graphs, Background Assumptions
  - Formal—Likelihood Methods

Data Analysis: Informal

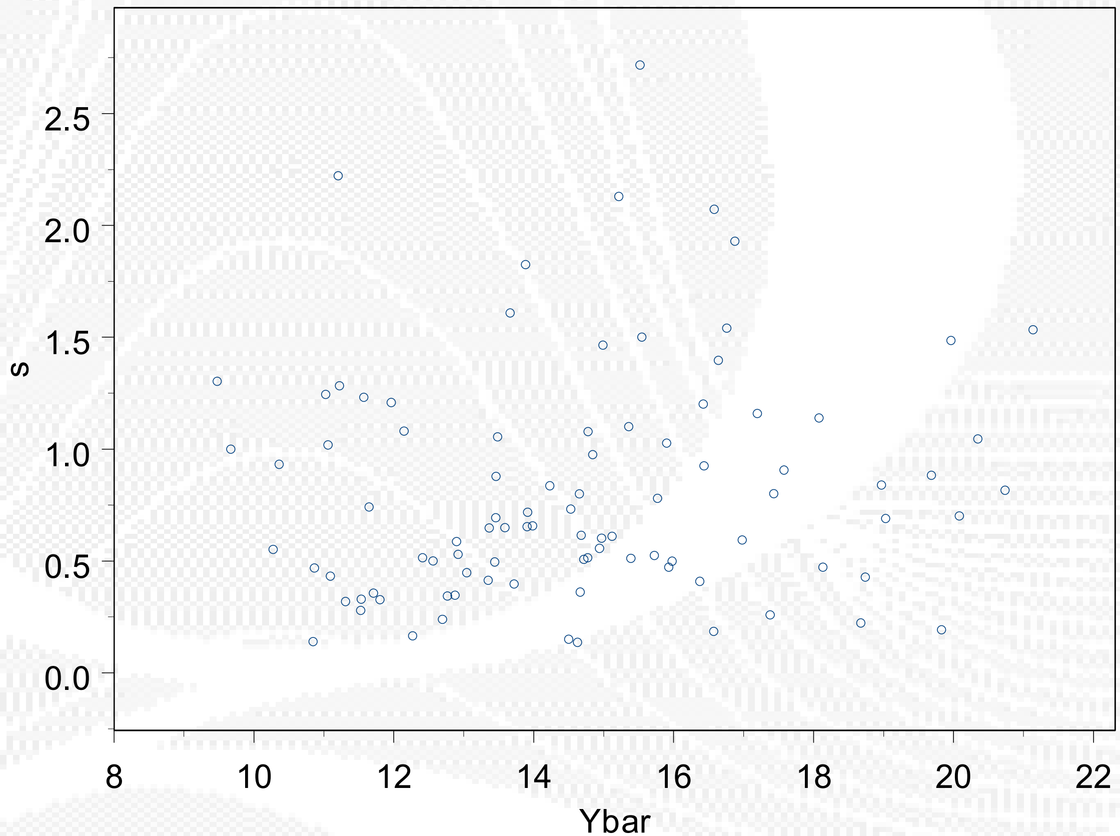
- The largest model we want to fit is Model 4.
  - But what if even *this* isn't right?
  - Can the data tell us?
- Yes, up to a point. Examples of informal analysis:
  - Does measurement variability increase as the values increase?
  - Is there a trend in three consecutive readings?
  - Is the bias, if any, linear?

Does Measurement Variability Increase as the Values Increase?

- Consider MD_y only here…

Plot of $s_{i,Y}$ vs. $\bar{Y}_{\bullet i}$

$$\boxed{\begin{array}{lcc} Y_{1i} & & \bar{Y}_{\bullet i} \\ Y_{2i} & \rightarrow & \\ Y_{3i} & & s_{i,Y} \end{array}}$$



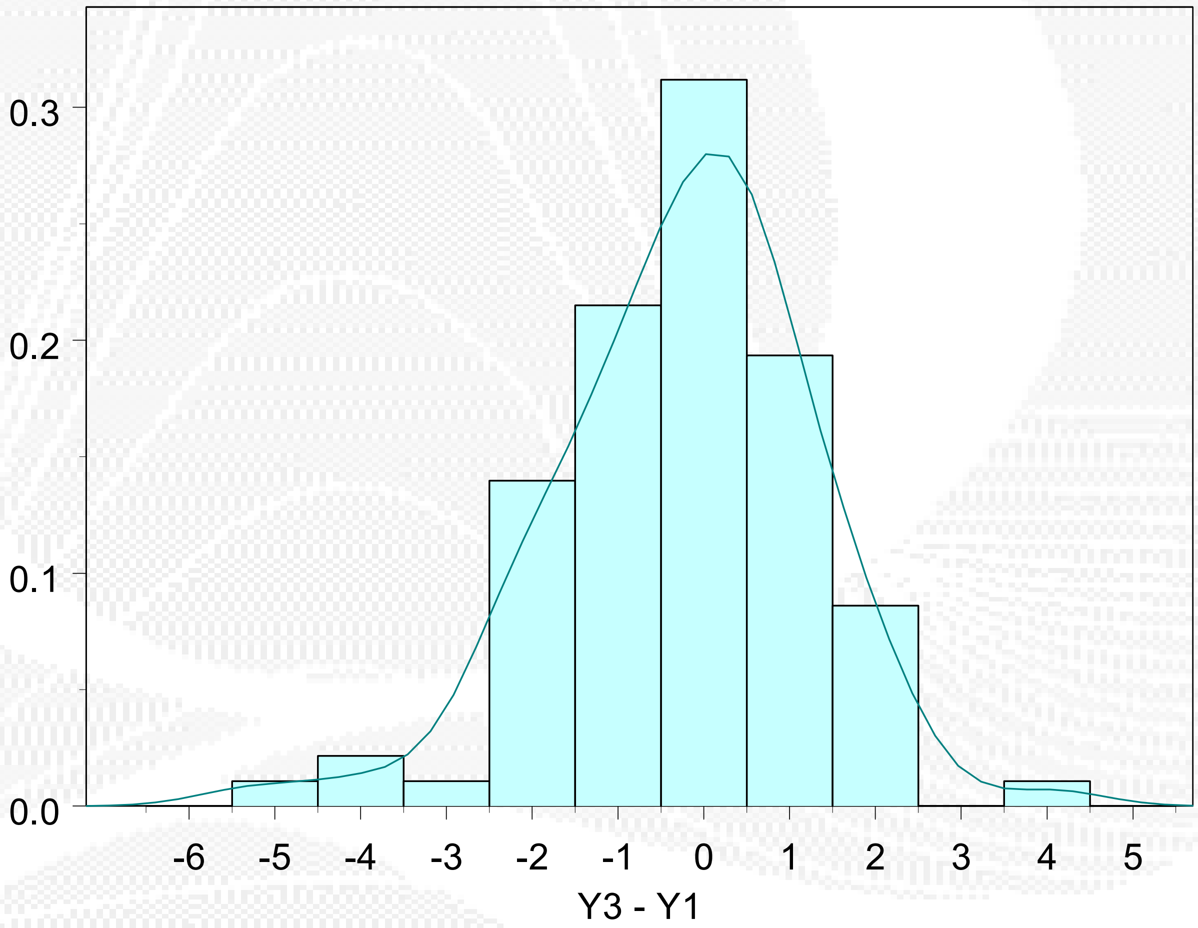
No evidence that $s_{i,Y}$ increases with $\bar{Y}_{\bullet i}$

Is there a Trend in Three Consecutive Readings?

Look at $Y_{3,i} - Y_{1,i}$

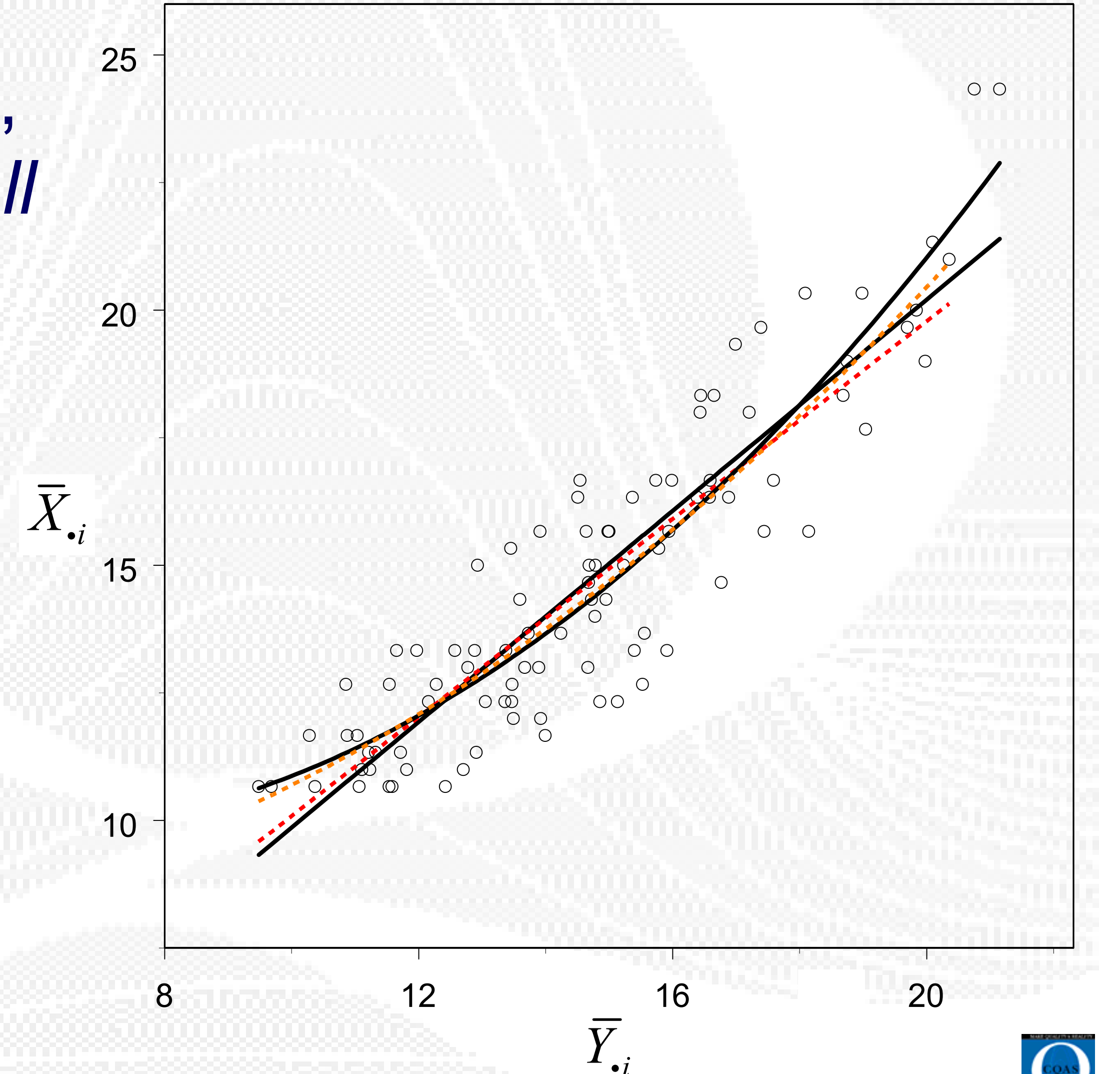
$$\boxed{\begin{array}{l} Y_{11} \\ Y_{21} \rightarrow Y_{31} - Y_{11} \\ Y_{31} \end{array}}$$

No evidence of a trend



Is the Bias, if any, Linear?

- Solid lines: linear, quadratic fits to *all* the data (N=93)
- Dashed lines: linear, quadratic fits without two largest *X* values
- →Set aside two largest *X* values

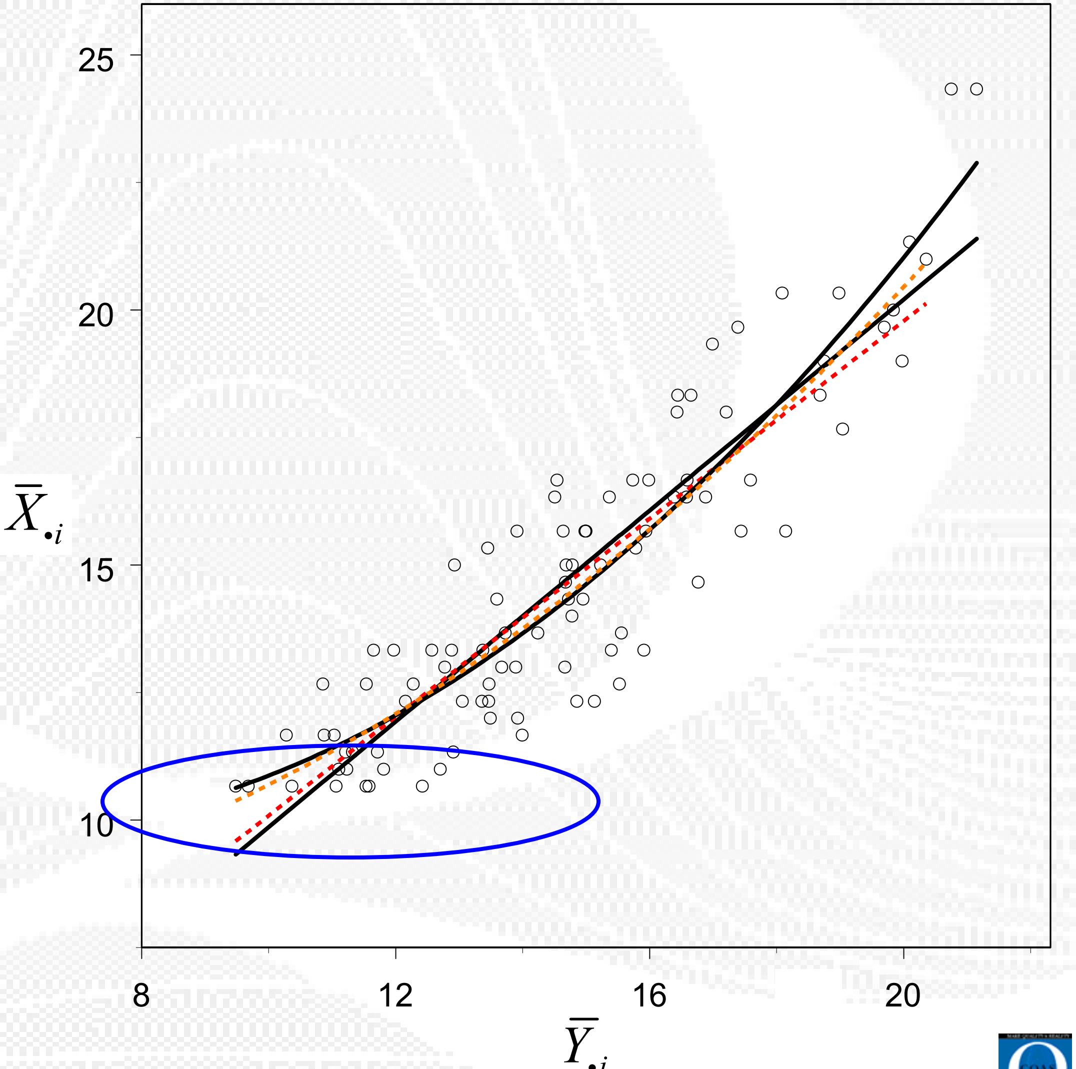


Another Lack of Fit?

Note "Boundary" of $\bar{X}_{\bullet i}$ at $\bar{X}_{\bullet i} \sim 10$

Set aside 7 lowest $\bar{X}_{\bullet i}$ values

Both high and low $\bar{X}_{\bullet i}$ features need to be investigated...



Topics

- Problem, Example, Mathematical Model
- Comparisons: Regression? Bland-Altman?
- More Models. Identifiability Problem, Bland-Altman
- A Richer Data Set and a Larger Model
- Comparison to Gage R&R
- Mandel's Estimates
- Data Analysis
  - Informal—Graphs, Background Assumptions
  - Formal—Likelihood Methods

CQAS

Data Analysis: Formal

- Comparison of Models in a hierarchy
- Start at largest and work down
- Find smallest model consistent with the data

Different

Model 4

Model 3

Model 2

Model 2′

Model 1

Lin Bias *and* Diff Prec

Lin Bias *or* Diff Prec

Equivalent

Data Analysis: Formal

- Estimation via *Maximum Likelihood*
- Compare models via *Likelihood Ratio Tests*
- Software? Coded in Excel, for client's needs.
  - Software via path diagrams available, e.g., Mx
  - Available in well-known statistical software??

	$\hat{\boldsymbol{\theta}}$ for Model k				
$\boldsymbol{\theta}$ for Model	$k=4$	$k=3$	$k=2$	$k=2'$	$k=1$
μ_x	14.805	14.806	14.806	14.834	14.824
β_0	1.248	0.504	1.653	0.000	0.000
β_1	0.918	0.968	0.891	1.000	1.000
σ_x^2	6.432	6.153	6.690	5.849	5.971
σ_δ^2	0.370	0.000	0.000	0.000	0.000
σ_e^2	3.119	3.398	2.115	3.422	2.139
σ_u^2	0.910	0.933	2.115	0.927	2.139
$(\mu_y = \beta_0 + \beta_1 \mu_x)$	14.842	14.842	14.842	14.834	14.824
$-2L(\hat{\boldsymbol{\theta}})$	1062.70	1065.01	1124.45	1065.41	1128.97
$-2L(\hat{\boldsymbol{\theta}})$ Difference, test of Model k versus $k-1$	2.31	59.44 vs 2 0.40 vs 2′	4.52	63.56	

$3.84/5.99$ critical value, $\alpha = 0.05$

Conclusions

- MD_x
  - Some unusual behavior at lowest and highest readings
  - Round-off error (seen in individual values).
- MD_y vs MD_x
  - Both MD's are measuring the same feature
  - No evidence of linear bias
  - MD_y is 1.9x more precise than MD_x
- Bland-Altman w/o reps: → lack of agreement?
  - But MD_y test, MD_x reference → wrong conclusions

Final Thoughts

- Structural models are natural models to use when comparing devices in the situation described in this talk
- Large literature, but not practiced much/well
  - Common technique such as regression, and the “recommended” method of Bland-Altman, can be misleading and so should be avoided
  - Software needs to be easily available
- Other modeling may be more appropriate to address other questions (such as operator consistency).